Japanese – Vietnamese compound noun translation

Vo Ho Bao Khanh
Graduate School of Media and Governance
Keio University
5322 Endo, Fujisawa, Kanagawa 252-8520, Japan
khanhvo@sfc.keio.ac.jp

Shun Ishizaki
Graduate School of Media and Governance
Keio University
5322 Endo, Fujisawa, Kanagawa 252-8520, Japan
ishizaki@sfc.keio.ac.jp

Abstract

Since compound nouns are especially popular and productive in Japanese, Japanese compound noun translation is an indispensable task. Many translation works of Japanese compound nouns to other popular languages have been done with the advantage of available resources such as available corpora and morphological processing tools. Nevertheless, there is no typical work of translating Japanese to other languages with less available resources despite the increasing needs of communication. Therefore, we would like to translate Japanese compound words to this kind of language, and choose Vietnamese as an illustrator because Japanese is very popular in Vietnam. In short, our proposed translation method consists of two processes: generating translation candidates and then selecting the correct candidates by using results of search engines.

1 Introduction

The translation of compound words is a major problem in not only multilingual dictionary compilation but also machine translation and multilingual information retrieval due to their frequency of occurrence and high productivity, especially in specialized documents (Fujii and Ishikawa, 1999). This translation of specialized compound words is very meaningful in expanding cutting-edge science and technology knowledge.

It is reported that most compound words are compound nouns, and the main pattern is noun-noun compound (Zhang and Isahara, 2004). In addition, compound noun is a popular phenomenon in oriental languages such as Japanese, Chinese and Vietnamese (Xin Zhao et al., 2006; Dinh Dien, 2002), thus many works of translating compound nouns from Japanese to other popular languages have been done.

On the other hand, the standard method that is often used to translate compound nouns consists of two basic stages: generation and selection (Tanaka and Baldwin, 2003). In generation stage, the compositional translation method is used to generate compound nouns into translation candidates. The most likely candidate then can be selected among these candidates by scoring the corpus-based translation quality of each candidate in monolingual corpus and cross-lingual data (Tanaka and Baldwin, 2004), or by applying the direct context-vector approach in comparable corpus (Morin and Daille, 2009). All these selection methods can be well performed on account of available revised corpora and fundamental language processing tools such as morphological processing tools, and part-of-speech tagger tools. Notwithstanding, such good corpora and language processing tools are not always available in some languages, especially in less popular languages. Briefly, another method is undoubtedly necessary for selecting and translating compound nouns. In this paper, we would like to propose such method. Our proposed method uses

results obtained from Web search engines in substitution of the mentioned selection approaches.

We decided to translate from Japanese compound nouns, chiefly noun-noun compounds, to Vietnamese because it can be treated as an illustration for less popular languages without available resources.

2 Compositional translation method

The compositional translation method has been widely used in translating compound words or multi-word terms (Morin and Daille, 2009). This method applies the composition of word-level translations, constructional translation templates and morphological property of constituents to generate translation candidates. The word-level translation can be performed easily if each constituent of a compound noun exists in bilingual dictionary. Translation templates can be constructed by grammatically investigating sample translation pairs. This compositional method has a weakness of depending much on the bilingual dictionary.

Typically, the word-level translation of Japanese compound nouns to Vietnamese made use of the general Japanese-Vietnamese dictionary consisting of about 60.000 word entries. The translation templates were compiled based on the detailed grammatical and semantic classification of Japanese compound nouns (Uchiyama et al., 2008). The translation templates then were verified by another compositional translation method that only analyzes the relationship of the source and target language compound nouns from the examples obtained from dictionaries or corpora. This method may miss out some translation templates if the investigated examples do not cover all constituent relations.

2.1 Grammatical classification of compound noun constituents

Each constituent of a compound noun is classified into different grammatical feature categories base on its part of speech information and its co-occurrence words (Uchiyama et al., 2008). Japanese compound noun constituents are classified into these following categories:

- Categories correspond to nouns: N1 and N2
- Categories correspond to verbal nouns: SN1 and SN2
- Adjective-like category: A
- Category corresponds with adjective and noun: N1A
- Other POS corresponding categories: E, G1, G2, G3

2.2 Grammatical and semantic relations of constituents

The above categorized constituents then are examined their relationships in compound nouns to set up common grammatical and semantic relationships between them. These relationships are divided into three main groups of ten concrete types with the assumption that we have two single nouns **α** and **β** of a compound noun **αβ**:

Table1: Grammatical and semantic relations of compound noun constituents

Set	Relations	Definition	Example
Case relation	α 'ga' β 'suru'	do β and α is a subject of β	人口集中
	α 'o' β 'suru'	do β with α is an object of β	要求分析
	α 'ni' β 'suru'	do β with α is an indirect object of β	文脈依存
Modification relation	α 'na' β, α 'teki na' β	α is a state of β	相対番地
	α 'ga' β 'de a ru' koto	α has an attribute of β	回復可能
	α 'de' β 'suru'	do β with or by α	機械翻訳
	α 'suru' ('tame no') β	β has a role of doing α	作業標準
	α 'ni' (α 'teki na') β 'suru'	do β with α is the state of doing β	自然結合
	α 'no' β	β of α	基本領域
Parallel or incidental relation	α 'shite' β 'suru', α 'shitara' β 'suru'	do α and β concurrently; or do α then do β	同期伝送, 予測解析

2.3 Translation template proposition

To construct detailed translation rules, we translated typical examples of each relation category into Vietnamese. The translation was made with the respect to the semantic relationships of Japanese compound nouns and Vietnamese language problems (Khanh Vo et al., 2009). We could propose 20 translation templates. Assuming that **γ** and **δ** are the correct translations of **β** and **α** respective-

ly, we have a sample of these translation templates in Table 2:

Table 2: Some translation template examples

Set	Relation	POS	Japanese	Vietnamese	Characteristics
Case relation	α 'ni' β 'suru'	α, β: verbal noun	α + β	γ + δ	Reversed order
	α 'ni' β 'suru'	α: noun, β: verbal noun	α + β	γ + preposition + δ	Reversed order and preposition
Modification	α 'de' β 'suru'	α: noun, expressing a state, β: verbal noun	α + β	γ + δ	Reversed order
	α 'de' β 'suru'	α: noun, an instrument, β: verbal noun	α + β	γ + preposition + δ	Reversed order and preposition
Parallel or incidental relation	α 'shite' β 'suru'	α: verbal noun, β: verbal noun	α + β	γ + (conjunction) + δ	Reversed order and conjunction

These proposed translation templates were verified again by example data extracted from the Japanese – Vietnamese dictionary and showed that all the proposed translation templates were sufficient.

3 Search engine-based selection method

After the compositional translation step, if we ignore the POS constraints, the number of generated translations is O(*mnt*) where *m* and *n* are the number of Vietnamese translations of each Japanese constituents and *t* is the number of translation templates. Accordingly, a large number of translations can be generated; hence the selection method is critical in determine the most likely translations. As mentioned in Section 1, due to the lack of available resources such as good corpora and fundamental language processing tools, we utilized Web search engines for our selection method.

Instead of searching for translation candidates in bilingual and comparable corpora, we searched them by available search engines in the Internet such as Bing, Google, and Yahoo!. We treated these candidates as keywords for searching. If a keyword exists in results of a search engine, we will get a number of those results and process them to eliminate incorrect results. For example, searching for the Vietnamese word "*tự kỷ chẩn đoán*" (self diagnosis) in Google can retrieve results containing punctuation marks such as "*tự kỷ, chẩn đoán*", "*tự kỷ; chẩn đoán*", or "*tự kỷ. Chẩn đoán*". We removed all the results containing these punctuation marks. We name this step as the pre-processing step. After the pre-processing step, we updated the number of results for each candidate up to its appropriate results. These results were sorted from the highest to the lowest. Only the results are greater than a threshold value were selected again to avoid the case that the highest hit is extremely greater than the lowest hit. The filtering threshold was calculated statistically basing the ratio of the really correct translations hits and the number of the retrieved results. Finally, we adopted two following selection methods for these filtered keywords and their search results:

3.1 The highest result selection

The highest result of each keyword for each search engine is selected from the sorted retrieved results. We then join these highest result keywords together to get the common keywords. If the common keywords exist, it means the most likely translation candidate has been selected.

3.2 Search engine confidence-based selection

To select the most likely candidates form search engines, the confidence value of each search engine needs to be determined. Typically, we need to determine **α, β, γ** in this linear formula:

$$\vec{X} = \alpha\vec{x_1} + \beta\vec{x_2} + \gamma\vec{x_3}$$

Since **α, β, γ** can be considered as the accuracy of each search engine ($0 \leq \alpha, \beta, \gamma \leq 1$), we can calculate these values based on the sample data and their retrieved results. This kind of value is the percentage of correct retrieved results and total retrieved results. For example, we determine the Google **α,** Yahoo! **β,** and Bing **γ** as 0.72, 0.81, and 0.64 respectively.

4 Experiment and evaluation

We collected 1000 compound nouns from different sources such as the patent documents provided by Industrial Property Digital Library of Japan and the Japanese – Vietnamese dictionary. These compound nouns are of a wide spectrum of specialized domains. First, these compound nouns were morphologically analyzed by Mecab (Kudo et al., 2004). Second, these compound nouns were translated compositionally with the support of Japanese-Vietnamese dictionary and translation templates. We could completely translate 768 compound nouns because the sparseness of the dictionary. The generated translation candidates of these compound nouns were selected by using both methods mentioned in Section 3. The highest result selec-

tion method gives the different results for individual search engines and the joined highest result. The search engine confidence-based selection method retrieves only a joined result of all search engines. Typically, the results of the highest result selection method are:

Table 3: The highest result selection method

Accuracy	Google Accuracy	Bing Accuracy	Yahoo! Accuracy
82.5%	82.8%	82.6%	82.9%

The search engine confidence-based selection method improved the result of the above method with the accuracy 85.4%.

5 Conclusion

This paper addresses the problem of translating compound nouns which are popular in a language to another language whose resources are not sufficient. The proposed method utilizing the compositional translation method and Web search results brought good results which can be a promising solution of not only the compound noun translation problem but also multiword translation problem.

References

Atsushi Fujii and Tetsuya Ishikawa. 1999. *Cross-Language Information Retrieval for Technical Documents.* In Proceedings of the 1999 Joint ACL SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-99), pp.29-37.Yujie Zhang, and Hitoshi Isahara. 2004. *Acquiring Compound Word Translations both Automatically and Dynamically.* In Proceedings of the 18th Pacific Asia Conference on Language, Information and Computation (PACLIC 18), pp.181-185.

Xin Zhao, Fuji Ren, and Shingo Kuroiwa. 2006. *Translation of Japanese Noun Compounds at Super-Function Based MT System.* In IEEJ Transactions on Electronics, Information and Systems, Vol. 126, No. 5, pp.645-653.

Dinh Dien. 2002. *Cognitive linguistics approach to Vietnamese noun compounds.* Mon-Khmer Studies: Journal of Southeast Asian Linguistics and Languages 32: pp.145-161.

Takaaki Tanaka and Timothy Baldwin. 2003. *Translation Selection for Japanese-English Noun-Noun Compounds.* In Proceedings of Machine Translation Summit IX, pp.378-85.

Timothy Baldwin and Takaaki Tanaka. 2004. *Translation by Machine of Complex Nominals: Getting it Right.* In Proceedings of the ACL 2004 Workshop on Multiword Expressions: Integrating Processing, pp. 24-31.

Emmanuel Morin and Béatrice Daille. 2009. *Compositionality and lexical alignment of multi-word terms.* In Journal of Language Resources and Evaluation, August, 2009.

Kiyoko Uchiyama, Shunsuke Aihara, Shun Ishizaki. 2008. *Identifying Semantic Relations in Japanese Compound Nouns for Patent Documents Analysis.* In Proceedings of Third International Conference on Large-Scale Knowledge Resources (LKR2008), pp. 75-81.

Khanh Ho Bao Vo, Kiyoko Uchiyama, Shun Ishizaki. 2009. *Applying grammatical feature analyses in multilingual compound noun translation.* In Proceedings of Pacific Association for Computational Linguistics, pp. 91-96.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. *Applying conditional random fields to Japanese morphological analysis.* In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pp. 230–237.